Towards a Logic of "Inferable" for Self-Aware Transparent Logical Agents

Stefania Costantini and Valentina Pitoni

DISIM, University of L'Aquila

Abstract. In Artificial Intelligence, multi agent systems constitute an interesting typology of society modeling, and have vast fields of application. Logic is often used to model such kind of systems as it is provides explainability and verifiabilty. In this paper[1] we talk about the cognitive aspect of autonomous systems, and we propose a particular logical framework (Logic of "Inferable" (*L-DINF*)) which introduces aspects of self-awareness, which is fundamental for reaching explainability. In fact, the proposed logic allow agents to reason about actions that they are able to perform, which is the logical inference chain that allows each action to be performed, and in how many steps. We consider resource-bounded agents, that can execute an action only if possessing the necessary resources to do so.

1 Introduction

According to the Oxford Handbook of Philosophy and Cognitive Science [9], Chapter by Alvin I. Goldman, *"Theory of Mind" refers to the cognitive capacity to attribute mental states to self and others. Other names for the same capacity include "commonsense psychology", "naïve psychology", "folk psychology", "mindreading" and "mentalizing". . . . In cognitive science the core question in this terrain is: How do people execute this cognitive capacity? How do they, or their cognitive systems, go about the task of forming beliefs or judgments about others' mental states, states that aren't directly observable? Less frequently discussed in psychology is the question of how people self-ascribe mental states.* In the literature, we can find different kinds of logical frameworks which try to emulate cognitive aspects of human beings. In this paper we propose a particular logical framework (Logic of "Inferable" (*L-DINF*)), in which we consider the concepts of step and of executability of an agent's actions.

Logics of 'awareness' have been studied in the recent years starting from the seminal work of Fagin & Halpern [10]. These logics distinguish between awareness, implicit belief and explicit belief. The crucial difference between our logic and existing logics of awareness is that the latter make no use of concepts as 'reasoning' or 'inference'. Instead, *L-DINF* provides a constructive theory of explicit beliefs, as it accounts for the perceptive and inferential steps leading from an agent's knowledge and beliefs to new beliefs; also, we considered two other important aspect: "steps" and "executability". The aspects related to epistemic attitudes is something our theory shares with other approaches in the literature including the dynamic theory of evidence-based beliefs by

[4], that also uses a neighborhood semantics for the notion of evidence, the sentential approach to explicit beliefs and their dynamics by [12], the dynamic theory of explicit and implicit beliefs by [17] and the dynamic logic of explicit beliefs and knowledge by [3].

The logic of inference stems from Velázquez-Quesada [16] and the logical system DES4_n by Duc [6], which share a similar view with our logic. In particular, Velázquez-Quesada shares with us the idea of modeling inference steps by means of dynamic operators in the style of dynamic epistemic logic (DEL). But, he does not emphasizes either the concepts of explicit belief and of background knowledge, nor issues related to the executability of actions. The system discussed in [6] shares with our logic the idea that an agent gets to know (or believe) something by performing inferences, and making inferences takes time. Nonetheless, while in our logic inferential operations are represented both at the syntactic level, via dynamic operators in the DEL style, and at the semantic level, as model-update operations, in Duc's system and its formal semantics they are not. In addition, we check whether an action can be performed or not and how many steps are needed to perform it.

Our constructive approach to explicit beliefs also distinguishes *L-INF* from Active logics [7,8], in which the basic semantics includes three components: (i) an agent's belief set, identifying all facts that an agent explicitly believes, (ii) an observation function, identifying all facts that an agent observes at a given time point, and (iii) an inference function, specifying what an agent should believe at the next time point on the basis of the application of her inference rules on her belief set, given her actual observations. There are important differences between active logics and our logic *L-INF*. Like active logics, we provide models of reasoning based on the assumption that an agent has a long-term memory and a short-term memory (or working memory). However, active logics do not distinguish, as we do, between the notion of explicit belief and the notion of background knowledge, conceived as different kinds of epistemic attitudes. In our approach, the long-term memory is 'stable', in the sense that it keeps the agent's background knowledge, that is assumed not to change during the agent's operation. The short-term memory instead contains beliefs (facts and rules) that either record the exogenous events that the agent has perceived or the new knowledge that it has learned, by its interactions with the external environment; it will also contain new beliefs that can be inferred, either locally or on the basis of rules retrieved from the long-term memory. In fact, our logic accounts for a variety of inferential actions (or 'mental operations') that have not been explored in the active logic literature and are very important to infer new beliefs. These actions correspond to basic operations of 'mind-reading' in the sense of Theory of Mind (ToM) [11], since they are mental and not physical ones, and require an agent to be aware of his own action's executability conditions, related to the available resources.

For our logic we we drew inspiration from the approach of *Self-aware computing*: quoting [15], *Self-aware and self-expressive computing describes an emerging paradigm for systems and applications that proactively gather information; maintain knowledge about their own internal states and environments; and then use this knowledge to reason about behaviors, revise self-imposed goals, and self-adapt.... Systems that gather unpredictable input data while responding and self-adapting in uncertain*

environments are transforming our relationship with and use of computers. Reporting from [1], *From an autonomous agent view, a self-aware system must have sensors, effectors, memory (including representation of state), conflict detection and handling, reasoning, learning, goal setting, and an explicit awareness of any assumptions. The system should be reactive, deliberative, and* **reflective**.

The introspective abilities of our agents as described in the proposed logic are limited. I.e., an agent is aware of the actions it can execute, and how many steps are needed to reach an objective. Nonetheless, our logic constitutes a first step towards self-aware agents.

The paper is organized as follows. In Section 2 we introduce the Syntax and the Semantic of our logic. In Sections 3 we show the axiomatization and the proof of soundness, instead in 4 we show the proof of strongl completeness. In Section 5 we propose a brief discussion on future work and conclude.

2 Logical framework

L-DINF is a logic which consists of a static component and a dynamic one. The static component, called *L-INF*, is a logic of explicit beliefs and background knowledge. The dynamic component, called *L-DINF*, extends the static one with dynamic operators capturing the consequences of the agents' inferential operations on their explicit beliefs as well as a dynamic operator capturing what an agent can conclude by performing some inferential operation in her repertoire.

2.1 Syntax

Assume a countable set of atomic propositions $Atm = \{p, q, \ldots\}$. By $Prop$ we denote the set of all propositional formulas, i.e. the set of all Boolean formulas built out of the set of atomic propositions Atm. A subset Atm_A of the atomic propositions represent the actions that an agent can perform.

The language of *L-DINF*, denoted by $\mathcal{L}_{L\text{-}DINF}$, is defined by the following grammar in Backus-Naur form:

$$\begin{array}{ll} \alpha & ::= \vdash(\varphi,\psi) \mid \cap(\varphi,\psi) \mid \downarrow(\varphi,\psi) \\ \phi,\psi & ::= p \mid exec(\alpha) \mid \neg\varphi \mid \varphi \wedge \psi \mid B\,\varphi \mid K\,\varphi \mid \\ & \quad [\alpha]\varphi \mid \Diamond\varphi \mid do(\phi_A) \mid can\_do(\phi_A) \end{array}$$

where the language of *inferential actions* of type α is denoted by $\mathcal{L}_{\mathsf{ACT}}$ (and a finite set of inferential actions X stems from $\mathcal{L}_{\mathsf{ACT}}$), and p ranges over Atm. Notice the expression $do(\phi_A)$, where it is required that $\phi_A \in Atm_A$. This expression indicates *actual execution* of action ϕ_A. In fact do is not axiomatized, as it is what has been called in [18] a *semantic attachment*, i.e., a procedure which connects an agent with its external environment in a way that is unknown at the logical level. $can\_do(\phi_A)$, where again $\phi_A \in Atm_A$, is a reserved syntax that, as seen later, will allow an agent to reason about its own capabilities.

Obviously the static part, *L-INF*, has the same definition save removing the inferential actions. The other Boolean constructions are defined from p, $\neg$ and $\wedge$ in the

standard way. The formula $B\,\varphi$ is read "the agent explicitly believes that φ is true" or, more shortly, "the agent believes that φ is true". Explicit beliefs are accessible in the working memory and are the basic elements of the agents' reasoning process, according the logic of local reasoning by Fagin & Halpern [10]. An effect of this approach is that agents cannot distinguish between logically equivalent formulas, i.e., if two facts φ and ψ are logically equivalent and an agent explicitly believes that φ is true, then she believes that ψ is true as well. There are other approaches, such as justification logics [14], that do not have this feature.

Unlike explicit beliefs, background knowledge is assumed to satisfy 'omniscience' principles like closure under conjunction and known implication, closure under logical consequence and introspection. Specifically, K is nothing but the well-known S5 operator for knowledge widely used in computer science. The fact that background knowledge is closed under logical consequence is justified by the fact that we conceive it as a kind of deductively closed *belief base*. Specifically, we assume background knowledge to include all facts that the agent has stored in her long-term memory (LTM), after having processed them in her working memory (WM), as well as all logical consequences of these facts.

The formula $[\alpha]\varphi$ should be read "φ holds after the inferential operation (or inferential action) α is performed by the agent". Borrowing from and extending [2], we include in the language of inferential actions $\mathcal{L}_{\mathsf{ACT}}$ three types of inferential operations α, which allow us to capture some of the dynamic properties of explicit beliefs and background knowledge. The operations that we consider are in particular the following: $\vdash(\varphi,\psi)$, $\cap(\varphi,\psi)$ and $\downarrow(\varphi,\psi)$. This operations characterize the basic operations of forming explicit beliefs via inference:

- $\vdash(\varphi,\psi)$ is the inferential operation which consists in inferring ψ from φ in case φ is believed and, according to an agent's working memory, ψ is a logical consequence of φ. With this last operation we operate directly on the working memory without retrieving anything from the background knowledge.
- $\downarrow(\varphi,\psi)$ is the inferential operation which consists in inferring ψ from φ in case φ is believed and, according to an agent's background knowledge, ψ is a logical consequence of φ. In other words, by performing this inferential operation, an agent tries to retrieve from her background knowledge in long-term memory the information that φ implies ψ and, if she succeeds, she starts to believe ψ; we assume that, differently from explicit beliefs, background knowledge is irrevocable in the sense of being stable over time.
- $\cap(\varphi,\psi)$ is the inferential operation which consists in closing the explicit belief that φ and the explicit belief that ψ under conjunction. In other words, $\cap(\varphi,\psi)$ characterizes the inferential operation of deducing $\varphi \wedge \psi$ from the explicit belief that φ and the explicit belief that ψ;

Remark 1. In the mental actions $\vdash(\varphi,\psi)$ and $\downarrow(\varphi,\psi)$, the formula ψ which is inferred and asserted as a new belief can be $do(\phi_A)$, which denotes the actual execution of action ϕ_A. In fact, we assume that when inferring $do(\phi_A)$ the action is actually executed, and the corresponding belief asserted, possibly augmented with a time-stamp. Actions are supposed to succeed by default, in case of failure a corresponding failure event will be

perceived by the agent. Therefore, the occurrence of physical actions (performed via the procedure do) is recorded through the formation of new beliefs.

The atomic formulas $exec(\alpha)$ have to be read "α is an inferential action that the agent can perform", meaning one of the above.

Remark 2. Specifying executability of actions pertains to the fact that we intend to consider agents with limited resources. Specific instances $\varphi\ OP_1\ \phi$ where one of the two formulas is an action may or may not be executable depending upon the agent having the necessary resources for performing that action. Executability is defined for inferential actions because such actions, if executable, may enact 'physical' actions.

Remark 3. Explainability in our approach can be directly obtained from proofs. If, for instance, the user asks an explanation of why the action ϕ_A has been performed, the system can respond by exhibiting the proof that has lead to ϕ_A, of the form $(exec(\varphi_1 OP_1\ \phi_1) \wedge (\varphi_1 OP_1\ \phi_1))\ \wedge \ldots\ \wedge (exec(\varphi_n OP_n\ \phi_A) \wedge \varphi_N OP_n\ do(\phi_A))$ where each OP_i is one of the (mental) actions discussed above. The proof can possibly be translated into natural language, and declined either top-down or bottom-up.

Finally, the formula $\Diamond\varphi$ has to be read "the agent can ensure φ by executing some inferential action in her repertoire". The interesting aspect of our language is that, at least in perspective (as the formalization is not complete yet), it allows us to express the concept of "being able to infer φ in k inferential steps". Specifically, let us inductively define: $\Diamond^0\varphi = \varphi$, $\Diamond^{k+1} = \Diamond\Diamond^k\varphi$. The formula $\Diamond^k B\,\varphi$ represents the fact that the agent is capable of inferring φ in k steps.

2.2 Semantics

The main semantics notions are outlined in the following definition of *L-INF* model which provides the basic components for the interpretation of the static version of the logic:

Definition 1. *We define a model to be a tuple $M = (W, N, R, E, V)$ where:*

- *W is a set of worlds or situations;*
- *$R \subseteq W \times W$ is an equivalence relation on W;*
- *$N : W \longrightarrow 2^{2^W}$ is a neighborhood function such that for all $i \in Agt$, $w, v \in W$ and $X \subseteq W$:*
  *(**C1**) if $X \in N(w)$ then $X \subseteq R(w)$,*
  *(**C2**) if wRv then $N(w) = N(v)$;*
- *$E : W \longrightarrow 2^{\mathcal{L}_{\mathsf{ACT}}}$ is an executability function such that for all $w, v \in W$:*
  *(**D1**) if wRv then $E(w) = E(v)$;*
- *$V : W \longrightarrow 2^{Atm}$ is a valuation function.*

For every $w \in W$, $R(w) = \{v \in W \mid wRv\}$ identifies the set of situations that the agent considers possible at world w. In cognitive terms, $R(w)$ can be conceived as the set of all situations that the agent *can* retrieve from her long-term memory and reason about them. More generally, $R(w)$ is called the agent's *epistemic state* at w.

For every $w \in W$, $N(w)$ defines the set of all facts that the agent explicitly believes at world w, a fact being identified with a set of worlds. More precisely, if $A \in N(w)$ then, at world w, the agent has the fact A under the focus of her attention and believes it. $N(w)$ is called the agent's explicit *belief set* at world w.

$E(w)$ is the set of mental operations that the agent can execute at w, as it has the resources to do so.

Constraint (**C1**) just means that an agent can have explicit in her mind only facts which are compatible with her current epistemic state. According to Constraint (**C2**), if world v is compatible with the agent's epistemic state at world w, then the agent should have the same explicit beliefs at w and v. Constraint (**D1**) means that an agent always knows the actions which she can perform and those that she cannot.

Truth conditions of *L-DINF* formulas are inductively defined as follows: for a model $M = (W, N, R, E, V)$, a world $w \in W$, a formula $\varphi \in \mathcal{L}_{L\text{-}INF}$, and an action α, we define the truth relation $M, w \models \varphi$ and a new model M^α by simultaneous recursion on α and φ as follows. Below, we write

$$\|\varphi\|_w^M = \{v \in W : wRv \text{ and } M, v \models \varphi\}$$

whenever $M, v \models \varphi$ is well-defined. Then, we set:

- $M, w \models p$ iff $p \in V(w)$
- $M, w \models exec(\alpha)$ iff $\alpha \in E(w)$
- $M, w \models \neg\varphi$ iff $M, w \not\models \varphi$
- $M, w \models \varphi \wedge \psi$ iff $M, w \models \varphi$ and $M, w \models \psi$
- $M, w \models B\,\varphi$ iff $||\varphi||_w^M \in N(w)$
- $M, w \models K\,\varphi$ iff $M, v \models \varphi$ for all $v \in R(w)$
- $M, w \models [\alpha]\varphi$ iff $M^\alpha, w \models \varphi$
- $M, w \models \Diamond\varphi$ iff $\exists\alpha \in E(w)$ s.t. $M^\alpha, w \models \varphi$

where $M^\alpha = (W, N^\alpha, R, E, V)$ and N^α is defined as follows. For $w \in W$, set

$$N^{\downarrow(\psi,\chi)}(w) = \begin{cases} N(w) \cup \{||\chi||_w^M\} & \text{if } M, w \models B\,\psi \wedge \\ & K\,(\psi \to \chi) \\ N(w) & \text{otherwise} \end{cases}$$

$$N^{\cap(\psi,\chi)}(w) = \begin{cases} N(w) \cup \{||\psi \wedge \chi||_w^M\} & \text{if } M, w \models B\,\psi \wedge \\ & B\,\chi \\ N(w) & \text{otherwise} \end{cases}$$

$$N^{\vdash(\psi,\chi)}(w) = \begin{cases} N(w) \cup \{||\chi||_w^M\} & \text{if } M, w \models B\,\psi \wedge \\ & B\,(\psi \to \chi) \\ N(w) & \text{otherwise} \end{cases}$$

We write $\models_{\text{L-DINF}} \varphi$ to denote that φ is true in all worlds w of every L-DINF model M.

Property 1. As consequence of previous definitions we have the following:

- $\models_{\text{L-INF}} (K(\varphi \to \psi)) \wedge B\,\varphi) \to [\downarrow(\varphi, \psi)]\, B\,\psi$.
  Namely, if an agent has φ as one of its beliefs and has $K(\varphi \to \psi)$ in its background knowledge, then as a consequence of the action $\downarrow(\varphi, \psi)$ the agent starts believing ψ;
- $\models_{\text{L-INF}} (B\varphi \wedge B\psi) \to [\cap(\varphi, \psi)]B_(\varphi \wedge \psi)$.
  Namely, if an agent has φ and ψ as beliefs, then as a consequence of the action $\cap(\varphi, \psi)$ the agent i starts believing $\varphi \wedge \psi$;
- $\models_{\text{L-INF}} (B(\varphi \to \psi)) \wedge B\,\varphi) \to [\vdash(\varphi, \psi)]\, B\,\psi$. Namely, if an agent has φ as one of its beliefs and has $B(\varphi \to \psi)$ in its working memory, then as a consequence of the action $\vdash(\varphi, \psi)$ the agent starts believing ψ;

3 Axiomatization

The *L-INF* and *L-DINF* axioms are:

1. $(K\,\varphi \wedge K(\varphi \to \psi)) \to K\,\psi$;
2. $K\,\varphi \to \varphi$;
3. $\neg K\,(\varphi \wedge \neg\varphi)$;
4. $K\,\varphi \to K\,K\,\varphi$;
5. $\neg K\,\varphi \to K\,\neg K\,\varphi$;
6. $B\,\varphi \wedge K\,(\varphi \leftrightarrow \psi) \to B\,\psi$;
7. $B\,\varphi \to K\,B\,\varphi$;
8. $\frac{\varphi}{K\,\varphi}$;
9. $[\alpha]p \leftrightarrow p$;
10. $[\alpha]\neg\varphi \leftrightarrow \neg[\alpha]\varphi$;
11. $exec(\alpha) \wedge [\alpha]\varphi \to \Diamond\varphi$;
12. $exec(\alpha) \to K\,(exec(\alpha))$;
13. $[\alpha](\varphi \wedge \psi) \leftrightarrow [\alpha]\varphi \wedge [\alpha]\psi$;
14. $[\alpha]K\,\varphi \leftrightarrow K\,([\alpha]\varphi)$;
15. $[\downarrow(\varphi, \psi)]B\,\chi \leftrightarrow B\,([\downarrow(\varphi, \psi)]\chi) \vee ((B\,\varphi \wedge K\,(\varphi \to \psi)) \wedge K\,([\downarrow(\varphi, \psi)]\chi \leftrightarrow \psi))$;
16. $[\cap(\varphi, \psi)]B\,\chi \leftrightarrow B\,([\cap(\varphi, \psi)]\chi) \vee ((B\,\varphi \wedge B\,\psi) \wedge K\,([\cap(\varphi, \psi)]\chi \leftrightarrow (\varphi \wedge \psi))$;
17. $[\vdash(\varphi, \psi)]B\,\chi \leftrightarrow B\,([\vdash(\varphi, \psi)]\chi) \vee ((B\,\varphi \wedge B\,(\varphi \to \psi)) \wedge B\,([\vdash(\varphi, \psi)]\chi \leftrightarrow \psi))$;
18. $\frac{\psi \leftrightarrow \chi}{\varphi \leftrightarrow \varphi[\psi/\chi]}$;
19. $p \to \Diamond p$;
20. $\Diamond(\varphi \wedge \psi) \to \Diamond\varphi \wedge \Diamond\psi$;
21. $\Diamond\varphi \to \Diamond\Diamond\varphi$;
22. $\Diamond B\,\varphi \to B\,\Diamond\varphi$;
23. $\Diamond K\,\varphi \to K\,\Diamond\varphi$.
24. $([\alpha]\varphi) \to \Diamond^1\varphi$.
25. $([\alpha_1]([\alpha_2]\varphi)) \to \Diamond^2\varphi$.
26. $([\alpha_1]([\alpha_2]([\alpha_3]\varphi))) \to \Diamond^3\varphi$

We write *L-DINF* $\vdash \varphi$ which signifies that φ is a theorem of $L\text{-}DINF$. Thanks to the previous axioms, *L-INF* and *L-DINF* are sound for the class of *L-INF* models.

This with the exception of the last axioms concerning $\Diamond^k$ that, in this formal setting, cannot be generalized, as we do not have the equivalence of "transitive closure". Therefore, in the present version one has to establish the maximum number of execution step to be considered (in the above formulation, just three steps) and write the corresponding axioms. These axioms are however of great practical utility for a self-aware agent: if the agent is able to prove $\Diamond^k \varphi$, where $\varphi = can\_do(\phi_A)$, then it becomes aware of being able to perform a certain action in k steps. As seen in the axioms, $\Diamond^k can\_do(\phi_A)$ is proved by finding a sequence of mental actions $([\alpha_1], \ldots [\alpha_k])$ which leads to the result. Therefore, the agent can possibly enact the analogous sequence (that, notice, is the equivalent of a *plan*), where $can\_do(\phi_A)$ is substituted by $do(\phi_A)$, so that action ϕ_A is eventually performed. This at the condition that all the $[\alpha_i]$'s are executable in the agent's present state.

4 Strong Completeness

For the proof that *L-INF* is strongly complete we limit ourselves, as we do not consider the $\Diamond$ operator, that is still an experimental feature that we aim to better formalize in future work. We can achieve the proof by means of a standard canonical model argument.

Definition 2. *The* canonical *L-INF* model *is a tuple* $M_c = \langle W_c; N_c; R_c; V_c; E_c \rangle$ *where:*

- W_c *is the set of all maximal consistent subsets of* $\mathcal{L}_{\textit{L-INF}}$*;*
- *For every* $w \in W_c$*,* wR_cv *if and only if* $K\,\varphi \in w$ *iff* $K\,\varphi \in v$*; i.e.,* R_c *is an equivalence relation on knowledge; as before, we define* $R_c(w) = \{v \in W_c \mid wR_cv\}$*.*
- *For* $w \in W_c$*,* $\varphi \in \mathcal{L}_{\textit{L-INF}}$ *we define* $A_\varphi(w) = \{v \in R_c(w) \mid \varphi \in v\}$*. Then, we put* $N_c(w) = \{A_\varphi(w) \mid B\,\varphi \in w\}$*.*
- V_c *is a valuation function defined as before;*
- E_C *is the executable function defined as before.*

There are the following immediate consequences of the above definition: if $w \in W_c$ then

- given $\varphi \in \mathcal{L}_{\text{L-INF}}$, it holds that $K\,\varphi \in w$ if and only if $\forall v \in W_c$ such that wR_cv we have $\varphi \in v$;
- for $\varphi \in \mathcal{L}_{\text{L-INF}}$, if $B\,\varphi \in w$ and wR_cv then $B\,\varphi \in v$;
- for $\alpha \in E_c(w)$, if wR_cv then $\alpha \in E_c(v)$.

Thus, R_{c_i}-related worlds have the same knowledge and N_c-related worlds have the same beliefs, i.e. there can be R_{c_i}-related worlds with different beliefs. The following results hold:

Lemma 1. *For all* $w \in W_c$ *and* $B\,\varphi, B\,\psi \in \mathcal{L}_{\textit{L-INF}}$*, if* $B\,\varphi \in w$ *but* $B\,\psi \notin w$*, it follows that there exists* $v \in R_c(w)$ *such that* $\varphi \in v \leftrightarrow \psi \notin v$*.*

Proof. Let $w \in W_c$ and φ, ψ be such that $B\,\varphi \in w$ and $B_i\,\psi \notin w$. Assume now that for every $v \in R_c(w)$ we have $\varphi \in v$, $\psi \in v$ or $\varphi \notin v$, $\psi \notin v$; then, from previous statements it follows that $K(\varphi \leftrightarrow \psi) \in w$ so that by axiom 6 in (3) $B\,\psi \in w$ which is a contradiction.

Lemma 2. *For all* $\varphi \in \mathcal{L}_{L-INF}$ *and* $w \in W_c$ *it holds that* $\varphi \in w$ *if and only if* $M_c, w \vDash \varphi$.

Proof. We have to prove the statement for all $\varphi \in \mathcal{L}_{\textit{L-INF}}$.

- $\varphi = p$, $w \in W_c$, if $p \in w$ then $p \in V_c(w)$ so for the truth conditions in definition (1) we have $M_c, w \vDash p$; to prove the opposite implication we have to proceed with the same reasoning;
- $\varphi = exec(\alpha)$, $w \in W_c$, if $exec(\alpha) \in w$ then $\alpha \in E_c(w)$ so for the truth conditions in definition (1) we have $M_c, w \vDash exec(\alpha)$;
- all the other cases have the same proof except $\varphi = B\,\varphi$. Assume $B\,\varphi \in w$ and $w \in W_c$:

$$A_\varphi(w) = \{v \in R_c(w) \mid \varphi \in v\} =$$
$$= \text{ by definition (1) } =$$
$$= \| \varphi \|_w^{M_c} \cap R_c(w)$$

so for the previous definition of canonical model:

$$N_c(w) = \{A_\varphi(w) \mid B\varphi \in w\}$$

than $\| \varphi \|_w^{M_c} \in N_c(w)$ and for definition (1) $M_c, w \vDash B\,\varphi$.
Suppose $B\,\varphi \notin w$ so $\neg B\,\varphi \in w$ and we have to prove $\| \varphi \|_w^{M_c} \cap R_c(w) \notin N_c(w)$.
Choose $A \in N_c(w)$, by definition we know $A = A_\psi(w)$ for some ψ with $B\,\psi \in w$.
By Lemma (1) there is some $v \in R_c(w)$ such that $\varphi \in v \leftrightarrow \phi \notin v$, so we have:

1. for ($\rightarrow$) thanks to the induction hp $v \in (\| \varphi \|_w^{M_c} \cap R_c(w)) \setminus A_\psi(w)$;
2. for ($\leftarrow$) thanks to the induction hp $v_I \in A_\psi(w) \setminus (\| \varphi \|_w^{M_c} \cap R_c(w))$;

than for (1) and (2) $A_\psi(w) \neq \| \varphi \|_w^{M_c} \cap R_c(w)$ and since $A = A_\psi(w)$ was arbitrary in $N_c(w)$ we conclude that $\| \varphi \|_w^{M_c} \cap R_c(w) \notin N_c(w)$, and so than $M_c, w \nvDash B\,\varphi$.

Lemma 3. *For all* $\varphi \in \mathcal{L}_{\textit{L-DINF}}$ *there exists* $\tilde{\varphi} \in \mathcal{L}_{\textit{L-INF}}$ *such that* $\mathrm{L-DINF} \vdash \Phi \leftrightarrow \tilde{\varphi}$ *(for any L-DINF formula we can find an equivalent L-INF formula).*

Proof. We have to prove the sentence for all $\varphi \in \mathcal{L}_{\textit{L-INF}}$ but we show the proof only for $\varphi = p$ because the others are proved analogously. By the axioms in Section (3) we have $[\alpha]p \leftrightarrow p$ and by axiom (18) we have $\dfrac{[\alpha]p \leftrightarrow p}{\varphi \leftrightarrow \varphi[[\alpha]p/p]}$ which means that we can replace $[\alpha]p$ with p in φ.

The previous lemmas allow us to prove the following theorems.

Theorem 1. L-INF *is strongly complete for the class of* L-INF *models.*

Proof. Any consistent set φ may be extended to a maximal consistent set of formulas $w^\star \in W_c$ and $M_c, w^\star \vDash \Phi$ by Lemma (2). Then, *L-INF* is strongly complete for the class of *L-INF* models.

Theorem 2. L-DINF *(without* $\Diamond$*) is strongly complete for the class of* L-INF *models.*

Proof. If K is a consistent set of $\mathcal{L}_{\textit{L-DINF}}$ formulas then $\tilde{K} = \{\tilde{\varphi} \mid \varphi \in K\}$ is a consistent set of $\mathcal{L}_{\textit{L-INF}}$ formulas by Lemma (3). By Theorem (1) there is a model M_c with a world w such that $M_c, w \vDash \tilde{K}$. But since *L-DINF* is sound and for each $\varphi \in K$, $\mathrm{L-DINF} \vdash \varphi \leftrightarrow \tilde{\varphi}$, it follows $M_c, w \vDash K$ then *L-DINF* is strongly complete for the class of *L-INF* models.

5 Conclusions

In this paper we discussed some cognitive aspects of autonomous system, concerning execution steps and executability. To model these aspects we have proposed the modal logic *L-INF*, and we have proved some useful properties among which strong completeness, though under significant restrictions.

In future work we mean to extend our logic to the multi-agents case and to prove, without restrictions, that *L-INF* is strongly complete. We will then extend *L-INF* to the multi-agents case and there is an intention to insert the concept of budget, as a value that represents how much an agent can spend of his own resources in the world w. This is important to better represent the fact that the agent is resource-bounded. Moreover, we have to compute the complexity and extend the group of inferential actions that we consider. Also, the temporal aspects of an agent's operation has not been considered here, but we tackled this aspects in [5,13]. We intende to merge the two approaches, so as to obtain a comprehensive framework.

References

1. Amir, E., Andreson, M.L., Chaudri, V.K.: Report on DARPA workshop on self aware computer systems. Technical Report, SRI International Menlo Park United States (2007)
2. Balbiani, P., Duque, D.F., Lorini, E.: A logical theory of belief dynamics for resource-bounded agents. In: Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, AAMAS 2016. pp. 644–652. ACM (2016)
3. Balbiani, P., Fernández-Duque, D., Lorini, E.: The dynamics of epistemic attitudes in resource-bounded agents. Studia Logica **107**(3), 457–488 (2019). https://doi.org/10.1007/s11225-018-9798-4
4. van Benthem, J., Pacuit, E.: Dynamic logics of evidence-based beliefs. Studia Logica **99**(1-3), 61–92 (2011). https://doi.org/10.1007/s11225-011-9347-x, https://doi.org/10.1007/s11225-011-9347-x
5. Costantini, S. and Pitoni, V.: A Temporal Module for Logical Frameworks. In: Proceedings 35th International Conference on Logic Programming (Technical Communications), ICLP 2019. EPTCS, Volume 306 (2019) 240–346
6. Duc, H.N.: Reasoning about rational, but not logically omniscient, agents. J. Log. Comput. **7**(5), 633–648 (1997). https://doi.org/10.1093/logcom/7.5.633
7. Elgot-Drapkin, J., Kraus, S., Miller, M., Nirkhe, M., Perlis, D.: Active logics: A unified formal approach to episodic reasoning (09 1996)
8. Elgot-Drapkin, J.J., Miller, M.I., Perlis, D.: Life on a desert island: Ongoing work on real-time reasoning (1987)
9. Eric Margolis, R.S., Stich(eds.), S.P.: The Oxford Handbook of Philosophy of Cognitive Science. Oxford University Press (2012). https://doi.org/10.1093/oxfordhb/9780195309799.001.0001
10. Fagin, R., Halpern, J.Y.: Belief, awareness, and limited reasoning. Artif. Intell. **34**(1), 39–76 (1987). https://doi.org/10.1016/0004-3702(87)90003-8, https://doi.org/10.1016/0004-3702(87)90003-8
11. Goldman, A.I., et al.: Theory of mind. The Oxford handbook of philosophy of cognitive science **1** (2012)
12. Jago, M.: Epistemic logic for rule-based agents. Journal of Logic, Language and Information **18**(1), 131–158 (2009). https://doi.org/10.1007/s10849-008-9071-8

13. Pitoni, V.: Memory Management in Resource-Bounded Agents. In: Proceedings 35th International Conference on Logic Programming (Technical Communications), ICLP 2019. EPTCS, Volume 306 (2019) 452–460
14. Savic, N., Studer, T.: Relevant justification logic. FLAP **6**(2), 397–412 (2019)
15. Tørresen, J., Plessl, C., Yao, X.: Self-aware and self-expressive systems. IEEE Computer **48**(7) (2015) 18–20
16. Velázquez-Quesada, F.R.: Explicit and implicit knowledge in neighbourhood models. In: Grossi, D., Roy, O., Huang, H. (eds.) Logic, Rationality, and Interaction - 4th International Workshop, LORI 2013, Hangzhou, China, October 9-12, 2013, Proceedings. pp. 239–252. Springer (2013). https://doi.org/10.1007/978-3-642-40948-6-19
17. Velázquez-Quesada, F.R.: Dynamic epistemic logic for implicit and explicit beliefs. Journal of Logic, Language and Information **23**(2), 107–140 (2014). https://doi.org/10.1007/s10849-014-9193-0
18. Richard W. Weyhrauch Prolegomena to a Theory of Mechanized Formal Reasoning Artificial Intelligence, 13(1-2) (1980) 133–170